The early maximum likelihood estimation model of audiovisual integration in speech perception

Andersen, Tobias

Published in:
Acoustical Society of America. Journal

Link to article, DOI:
10.1121/1.4916691

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

The early maximum likelihood estimation model of audiovisual integration in speech perception

Tobias S. Andersen[a)]
Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark

(Received 4 July 2014; revised 12 March 2015; accepted 17 March 2015)

Speech perception is facilitated by seeing the articulatory mouth movements of the talker. This is due to perceptual audiovisual integration, which also causes the McGurk–MacDonald illusion, and for which a comprehensive computational account is still lacking. Decades of research have largely focused on the fuzzy logical model of perception (FLMP), which provides excellent fits to experimental observations but also has been criticized for being too flexible, *post hoc* and difficult to interpret. The current study introduces the early maximum likelihood estimation (MLE) model of audiovisual integration to speech perception along with three model variations. In early MLE, integration is based on a continuous internal representation before categorization, which can make the model more parsimonious by imposing constraints that reflect experimental designs. The study also shows that cross-validation can evaluate models of audiovisual integration based on typical data sets taking both goodness-of-fit and model flexibility into account. All models were tested on a published data set previously used for testing the FLMP. Cross-validation favored the early MLE while more conventional error measures favored more complex models. This difference between conventional error measures and cross-validation was found to be indicative of over-fitting in more complex models such as the FLMP. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4916691]

[JFC] Pages: 2884–2891

I. INTRODUCTION

Speech perception is facilitated when the face of the talker is seen, as in face-to-face conversation, compared to when it is not, as in a phone conversation (Sumby and Pollack, 1954). This effect is stronger when auditory speech perception is poor and is an important aid for hearing impaired listeners (Grant *et al.*, 1998). The effect is widely believed to be caused by perceptual audiovisual integration rather than just a post-perceptual combination of information from auditory speech perception and lip-reading. The McGurk–MacDonald illusion is a striking demonstration of the perceptual nature of audiovisual integration of speech (MacDonald and McGurk, 1978; McGurk and MacDonald, 1976). In this illusion a speech sound, e.g., /ba/, is dubbed onto a video of a face articulating an incongruent phoneme, e.g., /ga/. This creates an illusory percept, in this example, of hearing /da/.

Decades of research on the computational mechanisms underlying audiovisual integration in speech perception has largely focused on the fuzzy logical model of perception (FLMP), which has been shown to provide good fits to empirical data in a number of studies (Massaro, 1998; Massaro and Cohen, 1983, 2000; Massaro *et al.*, 2011; Schwartz, 2010). The good fits of the FLMP have, however, been argued to be due to the model's flexibility rather than its ability to capture the underlying computational mechanisms (Andersen *et al.*, 2002; Cutting *et al.*, 1992; Myung and Pitt, 1997; Pitt, 1995; Pitt *et al.*, 2003; Schwartz, 2003, 2006; Vroomen and Gelder, 2000). Although much of this criticism has been addressed (Massaro, 2000, 2003; Massaro and Cohen, 1993; Massaro *et al.*, 2001) a consensus has not been reached despite the invocation of a broad spectrum of methods for model evaluation.

The current study has two main purposes. First, it introduces early maximum likelihood estimation (MLE) (Andersen *et al.*, 2005) as a new model of audiovisual integration in speech perception. Early MLE is based on the MLE model of multisensory integration of continuous representations (Ernst and Banks, 2002). By introducing a response boundary, as in signal detection theory (Green and Swets, 1966), the model can be applied to categorical responses. In this model, integration occurs before categorization, hence the name early MLE.

The idea of modeling audiovisual integration in speech perception based on a continuous internal representation is not new. The pre-labeling model introduced by Braida (1991) is also based on this idea but differs in the way it models the mechanism of audiovisual integration. In the pre-labeling model, auditory and visual internal representations are assumed to be orthogonal and integration occurs by basing the decision on the Pythagorean sum of the two. Thus, the model is inherently multidimensional. In addition to assigning separate dimensions to the perceptual modalities, the pre-labeling model can also assign multiple dimensions to the representation within modalities. A multidimensional phonetic representation is probably very realistic as speech perception relies on multiple perceptual features. This realism comes, however, at the cost of computational complexity because the multiple

[a)]Author to whom correspondence should be addressed. Electronic mail: toban@dtu.dk

 0001-4966/2015/137(5)/2884/8/$30.00

dimensions necessitate numerical evaluation of multidimensional integrals when fitting it. To avoid this problem, the current study is limited to models with a one-dimensional internal representation. The current study aims to show that this can be done without critical loss of realism when applying the models to data from experimental paradigms in which a single phonetic contrast is varied.

The MLE principle is also not new to models of audiovisual speech perception as it is inherent to the FLMP, which can be interpreted as MLE based on categorical representations (Massaro, 1998). Hence, integration happens after categorization in the FLMP and it can therefore be seen as a late MLE model. Although this difference between early and late MLE integration may seem subtle, the current study aims to show that there are great differences between the models in terms of parameterization and model complexity. These differences form the basis for the design of three related models all of which will be considered as alternatives to early MLE and the FLMP.

The other main purpose of this paper is to show that cross-validation effectively includes both goodness-of-fit and model flexibility in model evaluation, and provides meaningful selection of models. The development of methods for model evaluation in cognitive and perceptual science is an important field in its own right and the FLMP has had an important role in this field. Model evaluation methods that have been applied to the FLMP can be divided into three categories.

First, methods such as Akaike's information criterion (AIC; Akaike, 1974, Pitt *et al.,* 2002), Bayesian information criterion (Schwarz, 1978; Pitt *et al.,* 2002) and the root mean squared error (RMSE) corrected for the number of degrees of freedom (e.g., Massaro, 1998) depend on the goodness-of-fit penalized by a function of the number of free parameters. The problem with these methods is that since the number of free parameters is a poor measure of model complexity for non-linear models they do not always correct adequately for model complexity (Pitt *et al.*, 2002). This is problematic when evaluating the FLMP, which is non-linear (Myung and Pitt, 1997), especially in some regions of its parameter space (Andersen *et al.,* 2002; Schwartz, 2006).

Other methods, such as the Bayes factor (Massaro *et al.,* 2001; Myung and Pitt, 1997; Schwartz, 2006, 2010) and minimum description length (Pitt *et al.*, 2002) do not suffer from this problem but are algorithmically complex (Pitt *et al.*, 2002) although a simplifying assumption exists for the Bayes factor (see Schwartz, 2010).

Finally, cross-validation methods do not suffer from the same problems: They apply to all types of models and are straightforward to apply. They aim to estimate the generalization, or prediction, error, which is the expected error for new data not used in fitting the model parameters. The generalization error differs from the training error, the error for the data that were used in fitting the model parameters. Variability in data is generally due to fixed and random effects. Flexible models will, generally, fit closely to both types of variations. This is problematic because they have, so to speak, found a trend in randomness and this trend will, generally, not reappear in new data. This is called overfitting and causes flexible models to have high generalization errors. At the other end of the spectrum of complexity are models that are not sufficiently complex to capture the fixed effects. These models are said to under-fit and will have high training errors as well as high generalization errors. Somewhere between these two extremes lies the true model that fits the fixed effects perfectly. The true model will have higher training error than more flexible models because it cannot accommodate random variations in the data. This is why the training error is a poor criterion for evaluating models. The true model will, however, have the lowest possible generalization error, which is why the generalization error is the ideal criterion for evaluating models. The problem is that estimating the generalization error requires separate data for fitting the model and for evaluating the model. This increases the amount of data required. Pitt and Myung (2002) provide a good introduction to these concepts.

In cross-validation the data are split into a training set, which is used for fitting the model, and a test set, which is used for estimating the generalization error. The process of splitting, fitting, and evaluating is repeated so that all the data are used in the evaluation. In this way, cross-validation circumvents the requirement for separate training and test data. Each split is called a *fold* and the sum of the generalization error estimates across folds is called the test, or validation, error. The validation error is thus an estimate of the generalization error, which is based on the entire data set. Hastie *et al.* (2009) and MacKay (2003) provide introductions to cross-validation and compare it to other model evaluation techniques.

Data splitting can be done in several different ways: between observers, trials, conditions, or stimuli. It is important that the way that the data are split reflects how the model aims to generalize. The FLMP and other models of audiovisual integration aim at predicting the audiovisual percept based on the auditory and visual percepts, or, more generally, at generalizing perception across stimuli and modalities. Therefore, cross-validation splits should be made between stimuli within observers.

To ensure that models and methods are compared using representative data all model comparisons in the current study are based on the University of California Santa Cruz (UCSC) corpus (Massaro, 1998; Massaro *et al.,* 1995; Massaro *et al.,* 1993), which has been used extensively for comparing models of audiovisual integration of speech (Massaro, 1998; Massaro *et al.,* 2001; Schwartz, 2006, 2010; Wagenmakers *et al.,* 2004).

II. METHODS

A. Data

The data used in the current study are the UCSC corpus collected by Massaro and co-workers (Massaro, 1998; Massaro *et al.,* 1993; Massaro *et al.,* 1995) who kindly made it available online.[1] In this data set, 82 observers identified five auditory, five visual, and 25 audiovisual speech stimuli. The stimuli were synthesized using a speech synthesizer and an animated talking head. The auditory and visual stimuli were designed to fall approximately linearly on a continuum ranging from a clear /ba/ to a clear /da/. The audiovisual stimuli consisted of all the 25 possible combinations of the

auditory and visual stimuli. The observers identified the stimuli as /ba/ or /da/. All data were stored as the proportion of /da/-responses. According to the reports describing the experimental procedures, each stimulus was presented 24 times. Hence, multiplying the response proportion by 24 should yield the response counts, which should be integer values. This was not the case for several response proportions indicating that there was some variability in the number of stimulus presentations. This has prevented the usage of likelihood based error measures in the current study, which therefore uses error measures based on the squared error.

B. Models

1. Gaussian model without integration

The Gaussian model without integration only introduces a psychometric function in order to impose constraints based on the experimental design. The purpose of this model is two-fold. First, it is contained in some of the models of integration described here. Hence, comparing these models with the model without integration will provide a more detailed view of whether it is their mechanisms of integration or the psychometric function that determines their performance. Second, as the model without integration has the highest number of free parameters of the models in this study it will serve to show how the number of free parameters influences model performance in terms of goodness-of-fit and validation error differently. As such it serves as a baseline model with maximal complexity.

The psychometric function, $\Phi(S;\ c,\ \sigma)$, is here the Gaussian cumulative distribution function. It returns the probability of a /da/-response as a function of the stimulus level, $S = 1,\ldots,5$, where $S = 1$ indicates a clear /ba/ and $S = 5$ indicates a clear /da/. The psychometric function has two free parameters: the threshold parameter, c, denoting the 0.5 threshold and the standard deviation, σ, which determines the slope of the function. Hence, the psychometric function models the response proportions for five data points using two free parameters. For audiovisual stimuli, the stimulus level, $S_{AV} = 1,\ldots,5$, is determined by the stimulus level of the auditory component of the stimulus while the slope and threshold depend on the visual stimulus. Technically, this model can also be constructed so that the visual stimulus component determines the stimulus level while the slope and threshold depend on the auditory stimulus but, for simplicity, this model is not included in the current study. The complete model thus consists of seven psychometric functions: one auditory, one visual and five audiovisual. As each function contains two free parameters, the model has 14 free parameters. The way in which the visual stimulus influences auditory perception in this model does not reflect a perceptual integration process, which is why the model is referred to as a model without integration.

The psychometric function can be interpreted as a model of the underlying perceptual process (Gescheider, 1997). According to this model observers base their responses on a scalar internal representation value, x, of a stimulus feature that distinguishes /ba/ from /da/. If the value of the internal representation exceeds the threshold, c, the observer responds /da/. Otherwise the observer responds /ba/. The mapping of the stimulus onto the internal representation is stochastic due to additive noise. The values of x are thus distributed according to the normal probability density function, $\varphi(x;\ \mu,\ \sigma)$ with mean $\mu = S$. The probability of responding /da/ is the probability of x exceeding the threshold, $x > c$, which is given by the integral

$$\int_c^{\infty} \varphi(x;\mu,\sigma) = \Phi(\mu;c,\sigma) = \Phi(S;c,\sigma).$$

The psychometric function, thus allows us to transform response probabilities to probability densities on a continuous internal representation. This is of great interest because cross-modal integration of continuous internal representations of stimulus features such as spatial location or size has been successfully model by the MLE model (Alais and Burr, 2004; Ernst and Banks, 2002).

2. Early MLE

In the MLE model, the distributions, φ_A and φ_V, of the auditory and visual internal representation values, x_A and x_V, are assumed to be independent. Therefore, the maximum likelihood estimate of the corresponding audiovisual distribution φ_{AV} is the normalized product of the auditory and visual probability densities, φ_A and φ_V. This product is also a Gaussian distribution with a mean, μ_{AV}, which is a weighted sum of the means, μ_A and μ_V, of the distributions, φ_A and φ_V. The weights, w_A and w_V, are given by the expressions $w_A = r_A/(r_A + r_V)$ and $w_V = r_V/(r_A + r_V)$. Note that the weights are mutually dependent since $w_A = (1–w_V)$. The parameter, $r = \sigma^{-2}$, denotes the precision. The more precise, or reliable, modality is thus given greater weight. This is known as the information reliability principle and is in accordance with many observations in studies of multisensory perception (Andersen *et al.*, 2004; Alais and Burr, 2004; Ernst and Banks, 2002). The precision, r_{AV}, of the audiovisual distribution is given by the (unweighted) sum of the reliabilities, r_A and r_V, of the auditory and visual distributions. Hence, integration of information always leads to a more precise estimate according to the MLE model.

Inherent to MLE is the assumption that the auditory and visual internal representations are one and the same. Hence the threshold, c, should be the same for the auditory and visual internal representations, which it is not in the model without integration. Alignment of the representations and thresholds can be achieved by noticing that $\Phi(S;\ c,\ \sigma) = \Phi(S - c;\ 0,\ \sigma)$. This transformation has no effect on the psychometric function but it implies a shift of the mean, $\mu_A = S_A - c_A$ and $\mu_V = S_V - c_V$, of the probability density functions, φ_A and φ_V. This aligns the auditory, visual, and hence also the audiovisual internal representations so that the threshold is zero for all of them. It also contains an important constraint on the early MLE model: just as the stimulus levels, S_A and S_V, are fixed at integer values from 1 to 5, so are the means of the distributions within a modality evenly distributed with a distance of 1 between them. It

thus only takes two free parameters, c_A and c_V, to determine the means of the five auditory and five visual distributions. This is illustrated in Fig. 1. The early MLE model thus has four free parameters; two for the auditory and two for the visual psychometric function.

3. The weighted model

In early MLE, the reliability, r, of the auditory and visual modalities determines their weight, w. The weighted model releases this constraint and assigns a free parameter to the weight. The standard deviation of the audiovisual probability density function is given by summing of variances $\sigma_{av}^2 = w_a^2\sigma_a^2 + w_v^2\sigma_v^2$.

There are several reasons for why the weight given to each modality would not be determined (entirely) by its reliability. First, early MLE assumes that the distance between stimulus levels is identical for auditory and visual stimuli. If this assumption is violated the auditory and visual internal representations are scaled differently and it is not possible to determine the standard deviation of the auditory and visual probability densities relative to one another. This difference in scale will thus require an additional free parameter and it can be shown that adding this free parameter to the early MLE model makes it equivalent to the weighted model. Another reason is that stimuli in one modality may distract attention from the other modality. This could mean that the information in one modality is more reliable for unimodal stimuli, when attention is focused, than for bimodal stimuli, when attention is divided (Andersen *et al.*, 2005). The weighted model can take this effect into account.

4. The FLMP

In the FLMP, audiovisual integration is based on response probabilities (or, equivalently, fuzzy truth values). If P_a and P_v denote the auditory and visual response probability, respectively, then the audiovisual response probability is given by the normalized product of P_a and P_v,

$$P_{av} = \frac{P_a P_v}{P_a P_v + (1 - P_a)(1 - P_v)}.$$

Applied to the UCSC corpus, the FLMP requires 10 free parameters—five for the auditory response probabilities and five for the visual response probabilities.

Note that as the audiovisual probability distribution is based on the normalized product of the auditory and visual probability distributions, the FLMP can be interpreted as MLE based on a categorical internal representation (binomial in this case, multinomial in the general case of more than two response categories). Therefore, integration occurs after categorization and the FLMP can thus be considered as being based on late MLE.

5. Gaussian late MLE model

Early MLE has the potential advantage that the constraints imposed on the experimental design—using stimuli evenly spaced on a continuum—are incorporated into the model. The FLMP does not have this potential advantage. Any difference in the performance of the two models can thus be due to this as well as on the different ways (early vs late) they implement MLE. It is, however, possible to construct a late MLE model that contains a continuous internal representation. In this model, the auditory and visual response probabilities are calculated from the psychometric function exactly as in early MLE. The audiovisual response

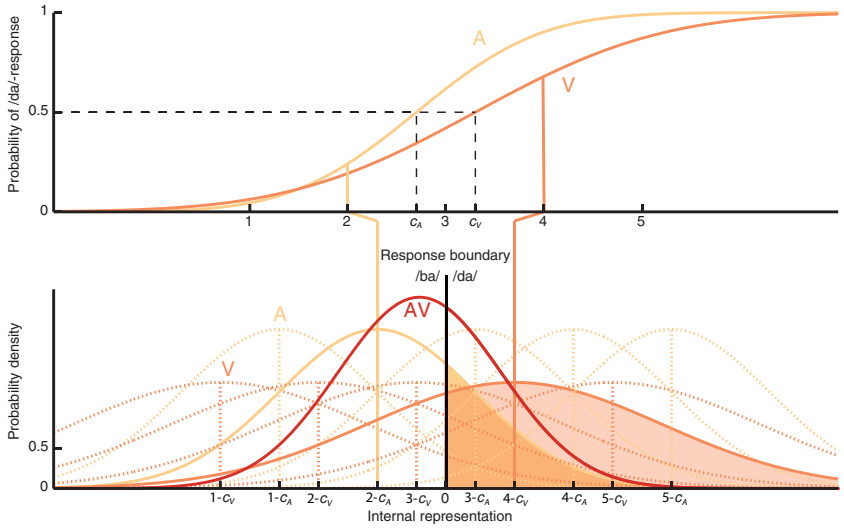

FIG. 1. (Color online) Illustration of the early MLE model. Upper axis: Example auditory (A) and visual (V) psychometric functions. Lower axis: Probability density functions of auditory (A) and visual (V) internal representation values corresponding to the psychometric functions in the upper axis. Each stimulus level in the upper axis determines the mean of a distribution in the lower axis. Examples are shown by lines connecting the axes. Note that the even spacing between stimulus levels in the upper axis is reflected in the even spacing between the distributions in the lower axis. The means of the five auditory and five visual distributions are thus determined by the auditory and visual thresholds, c_A and c_V, respectively. The example probability density function for the audiovisual internal representation values is calculated from MLE integration of the solid auditory and visual density functions. The response probability (of a /da/ response) is given by the probability mass falling above zero. Examples are shown by shaded areas.

probabilities are then calculated exactly as in the FLMP. This model is here termed the Gaussian late MLE model because it contains Gaussian noise in the early, continuous stage and late MLE as the model of integration. The parameters of the Gaussian late MLE model are the same as the parameters of the early MLE model.

6. Summary of models

The five models are summarized in Table I. The data set contains 35 data points (response proportions) for each subject (five auditory, five visual, and 25 audiovisual). In the five models there are three ways of reducing this complexity. First, the psychometric function (with no modeling of integration) reduces five degrees of freedom to two. Hence the Gaussian model without integration reduces 35 degrees of freedom to $2 \times 35/5 = 14$ free parameters. Second, modeling integration (without a psychometric function) predicts the 25 audiovisual data points from the five auditory and five visual data points. Hence the FLMP has $35 - 25 = 10$ free parameters. Finally, including both the psychometric function and a model of audiovisual integration predicts 35 data points from two psychometric functions. The early and late MLE models thus contain $2 + 2 = 4$ free parameters. The weighted model containing an additional free parameter for the weight contains five free parameters.

C. Fitting and cross-validation

The five models were all fitted to the data from each subject by minimizing the squared error between observed response proportions and the model response probabilities using the non-linear least squares solver from the Matlab™ Optimization Toolbox. As this is an unconstrained solver, constrained parameters were modeled as transformed unconstrained parameters. The weight, w, in the weighted model and response probabilities, P_a and P_v, in the FLMP were constrained to the range of 0 to 1 by applying a sigmoid function to unconstrained parameters. Standard deviations, σ, were constrained to be positive by applying the exponential function to unconstrained parameters.

Every model was fitted with 100 random initial conditions to minimize the chance of the optimization ending in a local minimum. The RMSE was calculated as the square root of the mean squared error for each subject. For each model, the RMSE corrected for degrees of freedom, henceforth referred to as the corrected RMSE, was calculated by dividing the RMSE by $(N_d - N_p)/N_d$, where N_d denotes the number of independent data points (35) and N_p denotes the number of free parameters.

Cross-validation was performed as a 35-fold leave-one-out procedure in which the models were fitted to the data from each subject separately. In each fold, the response proportion for one stimulus was left out from the fit. The validation squared error was then calculated between the model response probability and the observed response proportion for the stimulus left out from the fitting. The validation RMSE was then calculated as the square root of the across-fold mean squared error for each subject.

To test the significance of the differences in validation errors across models, the validation errors were subject to a one-way repeated measures analysis of variance (ANOVA). *Post hoc* tests were conducted in two ways. First, the

TABLE I. The parameters (pars.), their number (# pars.) and equations for the five models. P_A, P_V, and P_{AV} denote response probabilities for auditory, visual, and audiovisual stimuli, respectively. S_A, S_V, and S_{AV} denote stimulus level for auditory, visual, and audiovisual stimuli, respectively.

Model	Parameters	N_p	Description	Equations
Gaussian model w/o integration	C_A, σ_A C_V, σ_V C_{AV}, σ_{AV}	14	Thresholds and slopes for auditory, visual and five audiovisual psychometric functions	$P_a = \Phi(\mu_A;0, \sigma_A)$ $P_v = \Phi(\mu_V;0, \sigma_V)$ $P_{av} = \Phi(\mu_{AV};0, \sigma_{AV})$ $\mu_A = S_A - c_A$ $\mu_V = S_V - c_V$ $\mu_{AV} = S_{AV} - c_{AV}$
Early MLE	C_A, σ_A C_V, σ_V	4	Thresholds and slopes for auditory and visual psychometric functions	P_a, P_v, P_{av}, μ_A, and μ_V as in the Gaussian model w/o integration $\mu_{AV} = w_A\mu_A + w_V\mu_V$ $w_A = r_A/(r_A + r_V)$ $w_V = r_V/(r_A + r_V)$ $r_A = \sigma_A^{-2}$; $r_V = \sigma_V^{-2}$ $\sigma_{AV} = r_{AV}^{-0.5}$; $r_{AV} = r_A + r_V$
Weighted model	C_A, σ_A C_V, σ_V w_A	5	Thresholds and slopes for auditory and visual psychometric functions Weight parameters	P_a, P_v, P_{av}, μ_A, and μ_V as in the Gaussian model w/o integration $\mu_{AV} = w_A\mu_A + (1-w_A)\mu_V$ $\sigma_{AV} = \sqrt{w_a^2\sigma_a^2 + w_v^2\sigma_v^2}$
FLMP	P_a, P_v	10	Auditory and visual response probabilities	$P_{av} = \dfrac{P_aP_v}{P_aP_v + (1-P_a)(1-P_v)}$
Gaussian late MLE	C_A, σ_A C_V, σ_V	4	Thresholds and slopes for auditory and visual psychometric functions	P_a, P_v, μ_A, and μ_V as in the Gaussian model w/o integration P_{av} as in FLMP

validation error of each model was compared to every other model using a two-tailed *t*-test. Second, in another, less conventional, way, the models were ordered according to their validation error. Paired, one-sided *t*-tests were then performed between consecutive models. This was done in order to conduct *post hoc* tests with a smaller number of independent tests.

III. RESULTS

The results of the model fitting and cross-validation are displayed in Fig. 2 as the RMSE, the corrected RMSE and the validation RMSE. The models are ordered by number of free parameters so that models with more free parameters are to the left of models with fewer free parameters. The horizontal dashed line indicates the expectation value for the mean RMSE. The expectation value is calculated as the standard deviation of the response proportion assuming that the response count is distributed according to the binomial distribution with the response probability estimated by the observed response proportion.

As seen in Fig. 2, the differences in validation errors between models appear to be rather small. However, the ANOVA showed that the difference between the means of the validation errors is highly significant [$p < 0.001$, Greenhouse−Geisser corrected $F(2.6,\ 209.4) = 34.8$]. *Post hoc* paired two-tailed *t*-tests showed that the validation error of the early MLE model is significantly lower than the validation error of all of the other four models ($p < 0.0002$ for each comparison). The validation error of the weighted model is significantly lower than that of the late MLE, the FLMP and the early Gaussian model without integration ($p < 0.02$ for each comparison). The late MLE does not have lower validation error than the FLMP ($p > 0.9$) but both the late MLE and the FLMP has lower validation error than the Gaussian model without integration ($p < 10^{-6}$ for each comparison). When the models were ranked according to their validation error, paired one-sided *t*-tests, confirmed this pattern of significance. The *p*-values of these tests are displayed in Fig. 2.

The goodness-of-fit of over-fitting models is highly sensitive to small changes in parameter values. To test whether this is the case for the models described here, a sensitivity analysis was performed. For each subject a random number ranging from –5% to +5% of the parameter values was added to the best fitting unconstrained parameters. The RMSE was then calculated for these parameters. This procedure was repeated 1000 times and the mean difference between this RMSE and the RMSE of the best fit was calculated. This mean difference was small (<0.01) for all models compared to the difference in RMSE between models.

IV. DISCUSSION

The first purpose of the current study is to evaluate the early MLE model in comparison with the FLMP and the three other models described above. The early MLE model had the lowest validation error of the five models tested here and the difference in validation error between early MLE and the weighted model was highly significant. This finding shows that early MLE is a promising new model of audiovisual integration of speech.

However, this promise should be accompanied by words of caution. MLE models, early or late, contain a very strong constraint: the influence of each sensory modality depends only on the reliability of that modality. Audiovisual integration of speech may however vary across individuals (Magnotti and Beauchamp, 2014; Schwartz, 2010) beyond what can be explained due to variability in unimodal perception. Schwartz (2010) introduced a weighted version of the FLMP to account for this and showed that it performed better than the unweighted FLMP when applied to the UCSC corpus, the same data set as used here. The difference between Schwartz' findings and the findings in the current study may be due to differences in the type of weighted model and differences in the model evaluation methods. It is also possible that the individual differences in integration are distributed so that a majority of subjects integrate in agreement with early MLE while a significant minority integrates differently. Although early MLE performed significantly better than the weighted model in the current study, there was some variability across subjects and the weighted model was actually better for 24 out of 82 subjects.

Furthermore, several results in the literature suggest that audiovisual integration of speech can be influenced by the state of the observer without a corresponding change in unisensory perception (Alsius *et al.*, 2005; Nahorna *et al.*, 2012; Tuomainen *et al.*, 2005). These findings may require models

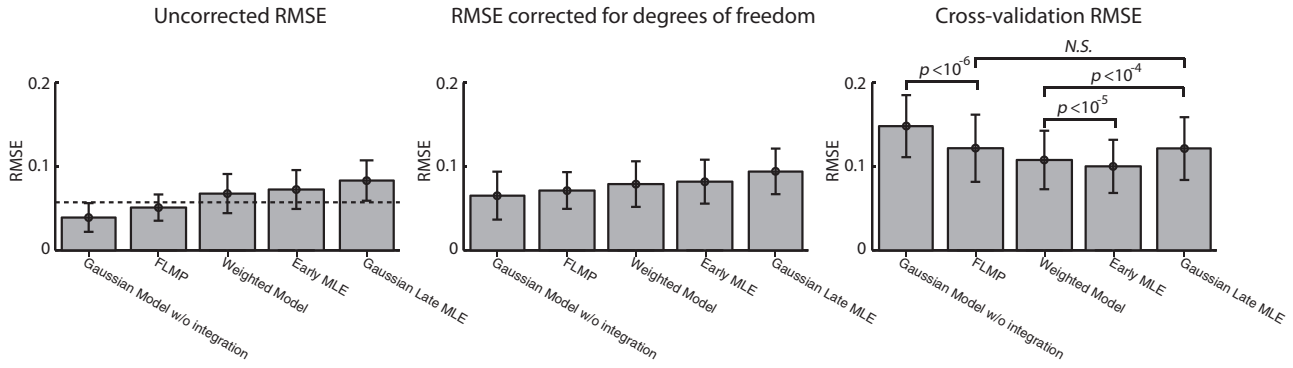

FIG. 2. The across-subject average RMSE, RMSE corrected for degrees of freedom, and validation RMSE for each of the seven models tested. Error bars represent the standard deviation (not the standard error of the mean as it would be too small to be clearly visible).

with a variable mechanism of integration such as weighted models or the Bayesian models of audiovisual integration suggested by Ernst (2006) and Shams *et al*. (2005).

So, how can the simple early MLE model perform well in the current study? The answer may lie in limitations in the data set. The single phonetic contrast, the limited number of stimuli along the stimulus continuum, the single signal-to-noise level and the lack of variation in the attentional state of the observer do not reflect the richness of everyday speech perception. Still, the data set has been influential and is not significantly smaller than data sets typically used in the literature on perceptual and cognitive models. This may indicate that the complexity of the data sets used to test models of audiovisual integration of speech so far has not matched the complexity of the models tested.

The weighted model and early MLE both had significantly lower validation error than Gaussian late MLE. This indicates that early integration reflects the mechanism of integration better than late MLE as this is the only difference between these two models.

Gaussian late MLE model did not have significantly lower validation error than the FLMP. This indicates that introducing the early continuous representation does not, in itself, lead to much improvement. This is confirmed by the FLMP having significantly lower validation error than the early Gaussian model without integration. From this we also learn that late MLE integration (Gaussian or FLMP) does seem to capture some of the underlying mechanism of integration, only not as well as early MLE.

The second purpose of the current study is to show that cross-validation effectively includes both goodness-of-fit and model flexibility in model evaluation, and provides meaningful selection of models. This is perhaps best seen by comparing model selection based on the validation error with model selection based on the corrected RMSE. Unsurprisingly, the RMSE consistently favored models with more free parameters. More importantly, this trend persisted when the RMSE was corrected for the degrees of freedom. Interestingly, this means that these measures did not favor the FLMP, in contrast to previous findings (Massaro, 1998), as the Gaussian model without integration, having the highest number of free parameters, had the lowest RMSE and corrected RMSE. This trend stands in stark contrast to the trend seen in the validation RMSE, which tends to favor the models with the fewest free parameters. The models with the more free parameters thus have low training errors and high validation errors, which is the hallmark of over-fitting. A further indication of over-fitting is that the RMSE was lower than the expectation value for the FLMP and the Gaussian model without integration. This suggests that these models fit not only to the variability due to fixed effects but also to variability due to the random effects.

The result of the sensitivity analysis indicated that all models were fairly robust to small variations in parameter values. Hence, although some models might over-fit in this study they do not do so to the extreme degree that was seen by Schwartz (2006) in a similar analysis of the FLMP. The reason for this discrepancy may be that Schwartz conducted his analysis on a different data set. This data set may have contained more response proportions close to zero for which the FLMP becomes highly non-linear and unstable.

That the early MLE is the best model of audiovisual integration of speech in terms of the cross-validation RMSE is a promising result. However, it may prove difficult to generalize it to more complex experimental designs that reflect real-life speech perception more closely. The reason for this is that whereas the continuous internal representation of speech is assumed to be one-dimensional in the current study, this is unlikely to be the case in general. Still, models with multidimensional representations do exist (Ashby, 1992) and it may be possible to insert a mechanism of integration into them. Although this may prove challenging, it also carries a promise: The inclusion of the experimental design in model design can lead to a more interpretable model with the dimensions of the model reflecting the perceptual features of audiovisual speech. Early MLE also contains a clear prediction for the effect of lowering the acoustic signal-to-noise ratio. This should lead to an increase in the variance of the Gaussian distribution in the auditory modality and increase the variability of responses across response categories as has been seen in early studies (Miller and Nicely, 1955). It should also lead to an increased visual influence in the McGurk illusion, which has also been reported (Sekiyama and Tohkura, 1991; Andersen *et al*., 2001). The FLMP can make no such prediction, as it does not parameterize the acoustic signal-to-noise ratio.

The conclusion of the current study is that cross-validation shows that audiovisual integration of speech is best modeled by the parsimonious early MLE model in the UCSD data corpus. Whether more complex models, such as multidimensional or weighted models, are required to model audiovisual integration of speech in general will require more complex data sets and is a task is left for future studies.

ACKNOWLEDGMENT

This study was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHESS).

[1]http://mambo.ucsc.edu/psl/8236/ (Last viewed June 6, 2010).

Akaike, H. (**1974**). "A new look at the statistical model identification," IEEE Trans. Autom. Control **19**, 716–723.

Alais, D., and Burr, D. (**2004**). "The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. **14**, 257–262.

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Curr. Biol. **15**, 839–843.

Andersen, T. S., Tiippana, K., Lampinen, J., and Sams, M. (**2001**). "Modeling of audiovisual speech perception in noise," in *International Conference on Auditory–Visual Speech Processing (AVSP)*, Aalborg, pp. 172–176.

Andersen, T. S., Tiippana, K., and Sams, M. (**2002**). "Using the fuzzy logical model of perception in measuring integration of audiovisual speech in humans," in *Proceedings of the First International NAISO Congress on Neuro Fuzzy Technologies*, Havana.

Andersen, T. S., Tiippana, K., and Sams, M. (**2004**). "Factors influencing audiovisual fission and fusion illusions," Brain Res. Cogn. Brain Res. **21**, 301–308.

Andersen, T. S., Tiippana, K., and Sams, M. (**2005**). "Maximum likelihood integration of rapid flashes and beeps," Neurosci. Lett. **380**, 155–160.

Ashby, F. G. (**1992**). "Multidimensional models of categorization," in *Multidimensional Models of Perception and Cognition*, edited by F. G. Ashby (Erlbaum, Hillsdale, NJ), pp. 449–483.
Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. A **43**, 647–677.
Cutting, J. E., Bruno, N., Brady, N. P., and Moore, C. (**1992**). "Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth," J. Exp. Psychol. **121**, 364–381.
Ernst, M. O. (**2006**). "A Bayesian view on multimodal cue integration," in *Human Body Perception From The Inside Out*, edited by G. Knoblich, I. M. Thornton, M. Grosjean, and M. Shiffrar (Oxford University Press, New York), Chap. 6, pp. 105–131.
Ernst, M. O., and Banks, M. S. (**2002**). "Humans integrate visual and haptic information in a statistically optimal fashion," Nature **415**, 429–433.
Gescheider, G. A. (**1997**). "Classical psychophysical theory," in *Psychophysics: The Fundamentals*, 3rd ed. (Psychology Press, East Sussex, UK), Chap. 4, pp. 73–103.
Grant, K. W., Walden, B. E., and Seitz, P. F. (**1998**). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," J. Acoust. Soc. Am **103**, 2677–2690.
Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).
Hastie, T., Tibshirani, R., and Friedman, J. (**2009**). "Model assessment and selection," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York), Chap. 7, pp. 219–257.
MacDonald, J., and McGurk, H. (**1978**). "Visual influences on speech perception processes," Percept. Psychophys. **24**, 253–257.
MacKay, D. (**2003**). "Model comparison and Occam's razor," in *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK), Chap. 28, pp. 343–353.
Magnotti, J. F., and Beauchamp, M. S. (**2014**). "The noisy encoding of disparity model of the McGurk effect," Psychonom. Bull. Rev. pp. 1–9.
Massaro, D. W. (**1998**). *Perceiving Talking Faces* (MIT Press, Cambridge, MA), 507 pp.
Massaro, D. W. (**2000**). "Reply to Vroomen and de Gelder," Trends Cognit. Sci. **4**, 38–39.
Massaro, D. W. (**2003**). "Model selection in AVSP: Some old and not so old news," in *International Conference on Auditory–Visual Speech Processing (AVSP)*, St. Jorioz, France, pp. 83–88.
Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol. **9**, 753–771.
Massaro, D. W., and Cohen, M. M. (**1993**). "The paradigm and the fuzzy logical model of perception are alive and well," J. Exp. Psychol. **122**, 115–124.
Massaro, D. W., and Cohen, M. M. (**2000**). "Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception," J. Acoust. Soc. Am. **108**, 784–789.
Massaro, D. W., Cohen, M. M., Campbell, C. S., and Rodriguez, T. (**2001**). "Bayes factor of model selection validates FLMP," Psychonom. Bull. Rev **8**, 1–17.
Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., and Tsuzaki, M. (**1993**). "Bimodal speech perception: An examination across languages," J. Phon. **21**, 445–478.
Massaro, D., Cohen, M. M., Meyer, H., Stribling, T., Sterling, C., and Vanderhyden, S. (**2011**). "Integration of facial and newly learned visual cues in speech perception," Am. J. Psychol. **124**, 341–354.
Massaro, D. W., Cohen, M. M., and Smeele, P. M. (**1995**). "Cross-linguistic comparisons in the integration of visual and auditory speech," Mem. Cognit. **23**, 113–131.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Myung, I. J., and Pitt, M. A. (**1997**). "Applying Occam's razor in modeling cognition: A Bayesian approach," Psychonom. Bull. Rev. **4**, 79–95.
Nahorna, O., Berthommier, F., and Schwartz, J.-L. (**2012**). "Binding and unbinding the auditory and visual streams in the McGurk effect," J. Acoust. Soc. Am. **132**, 1061–1077.
Pitt, M. A. (**1995**). "Data fitting and detection theory: Reply to Massaro and Oden," J. Exp. Psychol. **21**, 1065–1067 (1995).
Pitt, M. A., Kim, W., and Myung, I. J. (**2003**). "Flexibility versus generalizability in model selection," Psychonom. Bull. Rev. **10**, 29–44.
Pitt, M. A., and Myung, I. J. (**2002**). "When a good fit can be bad," Trends Cognit. Sci. **6**, 421–425.
Pitt, M. A., Myung, I. J., and Zhang, S. (**2002**). "Toward a method of selecting among computational models of cognition," Psychol. Rev. **109**, 472–491.
Schwarz, G. E. (**1978**). "Estimating the dimension of a model," Ann. Stat. **6**, 461–464.
Schwartz, J.-L. (**2003**). "Why the FLMP should not be applied to McGurk data or how to better compare models in the Bayesian framework," in *International Conference on Auditory–Visual Speech Processing (AVSP)*, St. Jorioz, France, pp. 77–82.
Schwartz, J.-L. (**2006**). "The 0/0 problem in the fuzzy-logical model of perception," J. Acoust. Soc. Am. **120**, 1795–1798.
Schwartz, J.-L. (**2010**). "A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent," J. Acoust. Soc. Am. **127**, 1584–1594.
Sekiyama, K., and Tohkura, Y. (**1991**). "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," J. Acoust. Soc. Am. **90**, 1797–1805.
Shams, L., Ma, W. J., and Beierholm, U. (**2005**). "Sound-induced flash illusion as an optimal percept," NeuroReport **16**, 1923–1927.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (**2005**). "Audio-visual speech perception is special," Cognition **96**, B13–B22.
Vroomen, J., and de Gelder, B. (**2000**). "Crossmodal integration: A good fit is no criterion," Trends Cognit. Sci. **4**, 37–38.
Wagenmakers, E.-J., Ratcliff, R., Gomez, P., and Iverson, G. J. (**2004**). "Assessing model mimicry using the parametric bootstrap," J. Math. Psychol. **48**, 28–50.